PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-081178
(43)Date of publication of application : 28.03.1997

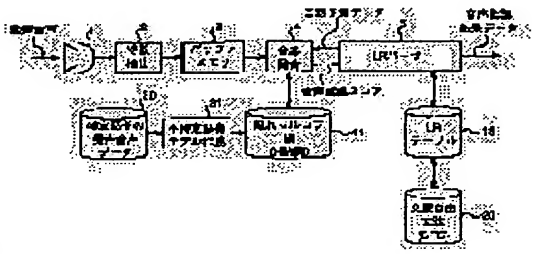
(51)Int.Cl. G10L 3/00
G10L 3/00

(21)Application number : 07-239821
(22)Date of filing : 19.09.1995

(71)Applicant : ATR ONSEI HONYAKU TSUSHIN KENKYUSHO:KK
(72)Inventor : TONOMURA MASAHIRO
MATSUNAGA SHOICHI

(54) UNSPECIFIED SPEAKER MODEL GENERATING DEVICE AND VOICE RECOGNITION DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To always generate an unspecified speaker model by independently clustering the output Gaussian distribution of each state of the hidden Markov model of single Gaussian distribution of plural specific speakers and synthesizing them.
SOLUTION: The unspecified speaker model generation section 31 learns the output Gaussian distribution only for the state, in which data exist, based on the uttered voice data of plurality N specific speakers stored in the memory of uttered voice data 30 of the specific speakers, extracts only the parameters of the learned output Gaussian distributions among the specific speaker's models and performs clustering for every state corresponding to the hidden Markov model(HMM). Then, synthesis and mixing are performed to generate a hidden Markov network (an HM network) 11 of the mixed Gaussian distributions and it is stored in the memory of the network 11. Then, voice recognition is performed by referring to the network 11. In other words, the output Gaussian distribution of each state of the hidden Markov model of a single Gaussian distribution of plural specific speakers is independently clustered for every state and an HMM is generated.



LEGAL STATUS

[Date of request for examination]	19.09.1995
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	2852210
[Date of registration]	13.11.1998

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平9－81178

(43)公開日　平成９年(1997)３月28日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ１０Ｌ　3/00	５２１		Ｇ１０Ｌ　3/00	５２１Ｎ	
				５２１Ｃ	
	５３５			５３５	

審査請求　有　　請求項の数５　ＯＬ　（全 10 頁）

(21)出願番号　特願平7－239821

(22)出願日　平成７年(1995)９月19日

(71)出願人　593118597
株式会社エイ・ティ・アール音声翻訳通信研究所
京都府相楽郡精華町大字乾谷小字三平谷５番地

(72)発明者　外村　政啓
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール音声翻訳通信研究所内

(72)発明者　松永　昭一
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール音声翻訳通信研究所内

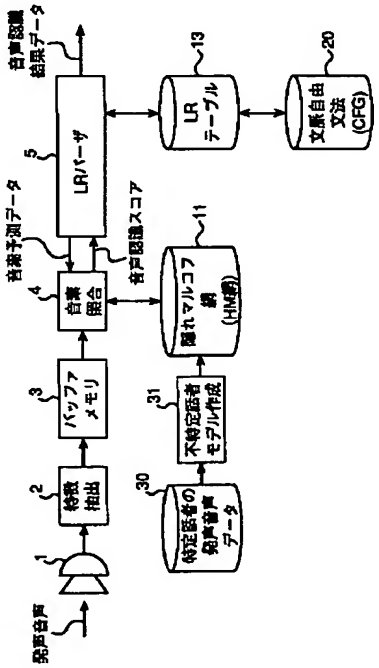
(74)代理人　弁理士　青山　葆　（外２名）

(54)【発明の名称】　不特定話者モデル作成装置及び音声認識装置

(57)【要約】

【課題】　各特定話者モデルのすべてのバラメータが学習されている必要がなく、また話者毎に学習されているバラメータが異なっている場合においても不特定話者モデルを作成できる不特定話者モデル作成装置及び音声認識装置を提供する。

【解決手段】　複数の特定話者の発声音声データに基づいて複数の話者に対して同一の初期話者モデルを用いてデータの存在する状態に対してのみ出力ガウス分布を学習することにより、複数個の特定話者用単一ガウス分布のＨＭＭを作成し、各出力ガウス分布間の距離を基準にして、各クラスタにより短い距離に出力ガウス分布が含まれるように複数のクラスタにクラスタリングを行う。次いで、各状態毎にクラスタリングされた単一ガウス分布のＨＭＭに基づいて、各クラスタ内の複数の出力ガウス分布のＨＭＭを各状態の単一ガウス分布のＨＭＭに合成した後混合することにより、不特定話者の混合ガウス分布のＨＭＭを作成する。



1

【特許請求の範囲】

【請求項１】　入力された複数の特定話者の単一ガウス分布の隠れマルコフモデルに基づいて、不特定話者の混合ガウス分布の隠れマルコフモデルを作成する不特定話者モデル作成装置において、

入力された複数の特定話者の単一ガウス分布の隠れマルコフモデルの各状態の出力ガウス分布を各状態ごとに独立にクラスタリングして合成することにより不特定話者の混合ガウス分布の隠れマルコフモデルを作成するモデル作成手段を備えたことを特徴とする不特定話者モデル作成装置。

【請求項２】　上記モデル作成手段は、

入力された複数の特定話者の発声音声データに基づいて、複数の話者に対して同一の初期話者隠れマルコフモデルを用いて所定の学習法により上記発声音声データの存在する状態に対してのみ出力ガウス分布を学習することにより、複数個の特定話者用単一ガウス分布の隠れマルコフモデルを作成する学習手段と、

上記学習手段によって作成された複数個の特定話者用単一ガウス分布の隠れマルコフモデルに基づいて、各出力ガウス分布間の距離を基準にして、各クラスタにより短い距離に出力ガウス分布が含まれるように複数のクラスタにクラスタリングを行うクラスタリング手段と、

上記クラスタリング手段によって各状態毎にクラスタリングされた単一ガウス分布の隠れマルコフモデルに基づいて、各クラスタ内の複数の出力ガウス分布の隠れマルコフモデルを各状態の単一ガウス分布の隠れマルコフモデルに合成する合成手段と、

上記合成手段によって合成された各状態の単一ガウス分布の隠れマルコフモデルを混合することにより、不特定話者の混合ガウス分布の隠れマルコフモデルを作成する混合手段とを備えたことを特徴とする請求項１記載の不特定話者モデル作成装置。

【請求項３】　上記クラスタリング手段は、各状態毎に予め設定したしきい値以上のデータ量で学習された出力ガウス分布のみを取り出した後、クラスタリングすることを特徴とする請求項２記載の不特定話者モデル作成装置。

【請求項４】　上記クラスタリング手段は、各状態においてクラスタリングされた各クラスタの中心と各出力ガウス分布間の距離の平均値が予め決めた距離以下になるまでクラスタリングを繰り返すことにより、各状態における各出力ガウス分布のバラツキが大きいほどクラスタ数が多くなるように各状態におけるクラスタ数を決定することを特徴とする請求項２又は３記載の不特定話者モデル作成装置。

【請求項５】　入力された複数の特定話者の単一ガウス分布の隠れマルコフモデルに基づいて、不特定話者の混合ガウス分布の隠れマルコフモデルを作成する請求項１乃至４のうちの１つに記載の不特定話者モデル作成装置

2

と、

入力された発声音声文の音声信号に基づいて、上記不特定話者モデル作成装置によって作成された不特定話者の混合分布の隠れマルコフモデルを用いて、音声認識する音声認識手段とを備えたことを特徴とする音声認識装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、複数の特定話者の隠れマルコフモデルに基づいて、不特定話者の隠れマルコフモデル（以下、ＨＭＭという。）を作成する不特定話者モデル作成装置、及びその不特定話者モデル作成装置を用いた音声認識装置に関する。

【０００２】

【従来の技術】従来、学習用の特定話者モデルに基づいて不特定話者のＨＭＭを作成するために、バーム・ウェルチ（Ｂａｕｍ－Ｗｅｌｃｈ）の学習アルゴリズム（以下、第１の従来例という。）が広く用いられている（例えば、中川聖一著，“確率モデルによる音声認識”，ｐｐ．５５－６４，電子情報通信学会，昭和６３年７月発行参照。）。この第１の従来例では、ＨＭＭにおいて時刻１から時刻ｔまでの間部分観測列｛y_1，y_2，y_3，…，y_t｝を観測した後、時刻ｔには状態ｉにいる前向き確率と、時刻ｔに状態ｉにいて時刻ｔ＋１から最後までの部分観測列｛y_{t+1}，y_{t+2}，y_{t+3}，…，y_r｝を観測する後向き確率とを用いて、ＨＭＭのパラメータを再推定して学習することにより、不特定話者のＨＭＭを作成する。

【０００３】上記第１の従来例の方法を用いて、多様な話者の音声の音響的特徴量の変動に対応するために多数話者の音声データでモデルを学習することが望ましく学習データが多量になる傾向があり、多数の話者による多量の音声データでモデルを学習することが望ましい。しかしながら、このような多量のデータを取り扱う場合、その膨大な計算量はコンピューターの処理速度が高速化しつつある現在においても問題となっている。

【０００４】このような不特定話者モデルの計算量を削減するために、既に小坂らによって特定話者モデルによる話者クラスタリングとモデル合成によるＣＣＬ法（以下、第２の従来例という。）が提案されている（従来文献２「小坂ほか，“クラスタリング手法を用いた不特定話者モデル作成法”，日本音響学会論文集，１－Ｒ－１２，１９９４年１１月」参照。）。この第２の従来例の方法では、各話者の音声の音響的特徴の類似性がすべての音響空間で等しいという仮定のもとに、すべての音韻にわたるモデルセット全体を単位としてクラスタリングを行っている。具体的には、十分に学習された特定話者モデルをモデル間の距離を定義することによってクラスタリングした後、各特定話者モデルを合成することにより不特定話者モデルを作成している。

【０００５】
【発明が解決しようとする課題】第２の従来例の方法では、少ない計算量で不特定話者モデルを作成することが可能であるが、特定話者モデルのすべてのパラメータが十分学習されていない場合には性能のよいモデルが得られないため各話者に対して多くの発声データが必要となる。また、HMMの全ての状態において混合出力ガウス分布の混合数が必ず同じになり、話者による特徴量のバラツキの少ない状態に対して無駄なパラメータが増えるという問題があった。
【０００６】本発明の第１の目的は以上の問題点を解決し、各特定話者モデルのすべてのパラメータが学習されている必要がなく、また話者毎に学習されているパラメータが異なっている場合においても不特定話者モデルを作成でき、しかも処理装置のメモリ容量が少なくてすみ、その計算時間を短縮することができる不特定話者モデル作成装置を提供することにある。また、本発明の第２の目的は、上記第１の目的に加えて、作成された不特定話者モデルを用いて音声認識することができ、従来例に比較して音声認識率を改善することができる音声認識装置を提供することにある。
【０００７】
【課題を解決するための手段】本発明に係る請求項１記載の不特定話者モデル作成装置は、入力された複数の特定話者の単一ガウス分布の隠れマルコフモデルに基づいて、不特定話者の混合ガウス分布の隠れマルコフモデルを作成する不特定話者モデル作成装置において、入力された複数の特定話者の単一ガウス分布の隠れマルコフモデルの各状態の出力ガウス分布を各状態ごとに独立にクラスタリングして合成することにより不特定話者の混合ガウス分布の隠れマルコフモデルを作成するモデル作成手段を備えたことを特徴とする。
【０００８】また、請求項２記載の不特定話者モデル作成装置は、請求項１記載の不特定話者モデル作成装置において、上記モデル作成手段は、入力された複数の特定話者の発声音声データに基づいて、複数の話者に対して同一の初期話者隠れマルコフモデルを用いて所定の学習法により上記発声音声データの存在する状態に対してのみ出力ガウス分布を学習することにより、複数個の特定話者用単一ガウス分布の隠れマルコフモデルを作成する学習手段と、上記学習手段によって作成された複数個の特定話者用単一ガウス分布の隠れマルコフモデルに基づいて、各出力ガウス分布間の距離を基準にして、各クラスタにより短い距離に出力ガウス分布が含まれるように複数のクラスタにクラスタリングを行うクラスタリング手段と、上記クラスタリング手段によって各状態毎にクラスタリングされた単一ガウス分布の隠れマルコフモデルに基づいて、各クラスタ内の複数の出力ガウス分布の隠れマルコフモデルを各状態の単一ガウス分布の隠れマルコフモデルに合成する合成手段と、上記合成手段によって合成された各状態の単一ガウス分布の隠れマルコフモデルを混合することにより、不特定話者の混合ガウス分布の隠れマルコフモデルを作成する混合手段とを備えたことを特徴とする。

【０００９】さらに、請求項３記載の不特定話者モデル作成装置は、請求項２記載の不特定話者モデル作成装置において、上記クラスタリング手段は、各状態毎に予め設定したしきい値以上のデータ量で学習された出力ガウス分布のみを取り出した後、クラスタリングすることを特徴とする。
【００１０】またさらに、請求項４記載の不特定話者モデル作成装置は、請求項２又は３記載の不特定話者モデル作成装置において、上記クラスタリング手段は、各状態においてクラスタリングされた各クラスタの中心と各出力ガウス分布間の距離の平均値が予め決めた距離以下になるまでクラスタリングを繰り返すことにより、各状態における各出力ガウス分布のバラツキが大きいほどクラスタ数が多くなるように各状態におけるクラスタ数を決定することを特徴とする。
【００１１】また、本発明に係る請求項５記載の音声認識装置は、入力された複数の特定話者の単一ガウス分布の隠れマルコフモデルに基づいて、不特定話者の混合ガウス分布の隠れマルコフモデルを作成する請求項１乃至４のうちの１つに記載の不特定話者モデル作成装置と、入力された発声音声文の音声信号に基づいて、上記不特定話者モデル作成装置によって作成された不特定話者の混合分布の隠れマルコフモデルを用いて、音声認識する音声認識手段とを備えたことを特徴とする。
【００１２】
【発明の実施の形態】以下、図面を参照して本発明に係る実施形態について説明する。図１は、本発明に係る一実施形態である音声認識装置のブロック図である。本実施形態の音声認識装置は、特に、特定話者の発声音声データ３０のメモリに格納された複数Ｎ人の特定話者の発声音声データに基づいて公知の最尤推定法を用いてデータの存在する状態に対してのみ出力ガウス分布を学習し、上記特定話者モデルの中から学習された出力ガウス分布のパラメータのみを取り出しHMMの対応する状態毎にクラスタリングを行った後合成及び混合を行って混合ガウス分布の隠れマルコフ網（以下、ＨＭ網という。）を作成し、作成したＨＭ網をＨＭ網１１のメモリに格納する不特定話者モデル作成部３１を備え、ＨＭ網１１のメモリに格納されたＨＭ網を参照して音声認識を行うことを特徴とする。
【００１３】この音声認識装置は、マイクロホン１と、特徴抽出部２と、バッファメモリ３と、音素照合部４と、文脈自由文法データベース２０のメモリに格納された所定の文脈自由文法に基づいて作成された、メモリに格納されたＬＲテーブル１３のメモリを参照して音声認識処理を実行する音素コンテキスト依存型ＬＲパーザ

5

（以下、ＬＲパーザという。）５とを備える。

【００１４】図２は、不特定話者モデル作成部３１によって実行される不特定話者モデル作成処理を示すフローチャートである。当該作成処理においては、まず、ステップＳ１において、複数Ｎ人の特定話者の発声音声データに基づいて、当該発声音声データの特徴パラメータを抽出し、抽出した特徴パラメータに基づいて、複数Ｎ人の全ての話者に対して同一のＨＭ網である初期話者モデル（各状態１混合）を用いて公知の最尤推定法によりデータの存在する状態に対してのみ出力ガウス分布の平均値と分散を学習することにより、Ｎ個の特定話者用単一ガウス分布のＨＭ網を作成する。

【００１５】次いで、ステップ２では、図３に示すように、作成されたＮ個の特定話者用単一ガウス分布のＨＭ網に基づいて、各状態毎に予め設定したしきい値以上のデータ量で学習された出力ガウス分布のみを取り出した後、図４に示すように、出力ガウス分布間の公知のバターチャ（Ｂｈａｔｔａｃｈａｒｙｙａ）距離を基準にして、各クラスタにより短い距離に出力ガウス分布が含まれるように複数のクラスタにクラスタリングを行なう。ここで、取り出す学習データ量にしきい値を設けたのは信頼性の低い出力ガウス分布がクラスタリングに悪影響を及ぼさないようにするためである。これにより、信頼性の高いＨＭ網１１を得ることができ、当該ＨＭ網１１を用いて音声認識することにより、従来例に比較して高い音声認識率で音声認識することができる。また、当該クラスタリングでは、各状態においてクラスタリングされ各クラスタの中心と各出力ガウス分布間の公知のバターチャ（Ｂｈａｔｔａｃｈａｒｙｙａ）距離の平均値が予め決めた距離以下になるまでクラスタリングを繰り返すことにより、各状態における各メンバーの出力ガウス分布のバラツキに応じてクラスタ数Ｋを決定する。ここで、バラツキが大きい場合はクラスタ数Ｋを比較的多く設定する一方、バラツキが小さい場合はクラスタ数Ｋを比較的少なく設定する。また、上記クラスタ数Ｋの決定においては、最大のクラスタ数Ｋｍａｘ及び最小のクラスタ数Ｋｍｉｎを設定してもよい。さらに、学習データ量が小さい場合は、好ましくは、クラスタ数Ｋを小さく設定する。

6

【００１６】次いで、ステップＳ３においては、上記ステップＳ２で各状態ごとにクラスタリングされた結果を用いて、図５に示すように、クラスタ内の複数の出力ガウス分布を各状態の単一ガウス分布に合成する。合成は出力ガウス分布の総数、及びクラスタリング結果が各状態ごとに異なること以外は、従来文献２の方法と同様の方法で行なった。当該ステップＳ３の合成方法については詳細後述する。さらに、ステップＳ４においては、各状態ごとに全てのクラスタの合成された単一ガウス分布を公知の話者混合法を用いて混合することにより混合ガウス分布のＨＭ網を作成してＨＭ網１１のメモリに格納する。混合比率は各クラスタのメンバーの出力ガウス分布の学習データ量の総和の比に比例する値とした。すなわち、各クラスタのメンバーの学習データ量が大きいほど、混合比率を大きく設定する。

【００１７】上記ステップＳ３において用いられる各クラスタにおける合成後の平均値μh_iと分散Sh_iは、次の数１及び数２で表される。なお、重み係数$w_i^{(1)}$は次の数３で表される。

【００１８】

【数１】

$$\mu h_i = \sum_j w_j^{(1)} \mu_j^{(1)}$$

【数２】

$$Sh_i = \sum_j w_j^{(1)} S_j^{(1)} + \sum_j w_j^{(1)} (\mu_j^{(1)} - \mu h_i)^2$$

【数３】

$$w_j^{(1)} = n_j^{(1)} / \{ \sum_i n_i^{(1)} \}$$

【００１９】数１と数２はそれぞれ、複数のガウス分布を単一ガウス分布と見なして求めた場合の平均値、分散を表す。ここで、$\mu_i^{(1)}$と$S_i^{(1)}$は自然数ｉ番目のＨＭ網のの状態ｊにおける単一ガウス分布である出力確率密度関数の平均値と分散を表わす。また、$n_i^{(1)}$はｉ番目のＨＭ網の状態ｊにおけるサンプル数を表す。すなわち、数１から明らかなように、合成後の平均値μh_iと分散Sh_iとはそれぞれ、合成前の平均値μ_iと分散S_iを、各状態におけるサンプル数$n_i^{(1)}$に応じてサンプル数$n_i^{(1)}$が大きいほど大きい重み係数$w_i^{(1)}$で重み付けされて計算される。

【００２０】本実施形態においては、音声認識のための統計的音素モデルセットとしてＨＭ網１１を使用している。当該ＨＭ網１１は効率的に表現された音素環境依存モデルである。１つのＨＭ網は多数の音素環境依存モデルを包含する。ＨＭ網１１はガウス分布を含む状態の結合で構成され、個々の音素環境依存モデル間で状態が共有される。このためパラメータ推定のためのデータ数が不足する場合も、頑健なモデルを作成することができる。このＨＭ網１１は逐次状態分割法（Successive State Splitting:以下、ＳＳＳという。）を用いて自動作成される。上記ＳＳＳではＨＭ網のトポロジーの決定、異音クラスタの決定、各々の状態におけるガウス分布のパラメータの推定を同時に行なう。本実施形態においては、ＨＭ網のパラメータとして、ガウス分布で表現される出力確率及び遷移確率を有する。このため認識時には一般のＨＭＭと同様に扱うことができる。

【００２１】次いで、上述の本実施形態の音声認識方法を用いた、ＳＳＳ－ＬＲ（left-to-right rightmost

型）不特定話者連続音声認識装置について説明する。この装置は、メモリに格納されたＨＭ網１１と呼ばれる音素環境依存型の効率のよいＨＭＭの表現形式を用いている。また、上記ＳＳＳにおいては、音素の特徴空間上に割り当てられた確率的定常信号源（状態）の間の確率的な遷移により音声パラメータの時間的な推移を表現した確率モデルに対して、尤度最大化の基準に基づいて個々の状態をコンテキスト方向又は時間方向へ分割するという操作を繰り返すことによって、モデルの精密化を逐次的に実行する。

【００２２】図１において、話者の発声音声はマイクロホン１に入力されて音声信号に変換された後、特徴抽出部２に入力される。特徴抽出部２は、入力された音声信号をＡ／Ｄ変換した後、例えばＬＰＣ分析を実行し、対数パワー、１６次ケプストラム係数、Δ対数パワー及び１６次Δケプストラム係数を含む３４次元の特徴パラメータを抽出する。抽出された特徴パラメータの時系列はバッファメモリ３を介して音素照合部４に入力される。

【００２３】音素照合部４に接続されるメモリ内のＨＭ網１１は、各状態をノードとする複数のネットワークとして表され、各状態はそれぞれ以下の情報を有する。

（ａ）状態番号
（ｂ）受理可能なコンテキストクラスタ
（ｃ）先行状態、及び後続状態のリスト
（ｄ）出力確率密度分布のパラメータ
（ｅ）自己遷移確率及び後続状態への遷移確率

【００２４】音素照合部４は、音素コンテキスト依存型ＬＲパーザ５からの音素照合要求に応じて音素照合処理を実行する。そして、不特定話者モデルを用いて音素照合区間内のデータに対する尤度が計算され、この尤度の値が音素照合スコアとしてＬＲパーザ５に返される。このときに用いられるモデルは、ＨＭＭと等価であるために、尤度の計算には通常のＨＭＭで用いられている前向きパスアルゴリズムをそのまま使用する。

【００２５】一方、メモリ内の所定の文脈自由文法（ＣＦＧ）データベース２０を公知の通り自動的に変換してＬＲテーブルを作成してＬＲテーブル１３のメモリに格納される。ＬＲパーザ５は、上記ＬＲテーブル１３を参照して、入力された音素予測データについて左から右方向に、後戻りなしに処理する。構文的にあいまいさがある場合は、スタックを分割してすべての候補の解析が平行して処理される。ＬＲパーザ５は、ＬＲテーブル１３から次にくる音素を予測して音素予測データを音素照合部４に出力する。これに応答して、音素照合部４は、その音素に対応するＨＭ網１１内の情報を参照して照合し、その尤度を音声認識スコアとしてＬＲパーザ５に戻し、順次音素を連接していくことにより、連続音声の認識を行い、その音声認識結果データを出力する。上記連続音声の認識において、複数の音素が予測された場合は、これらすべての存在をチェックし、ビームサーチの方法により、部分的な音声認識の尤度の高い部分木を残すという枝刈りを行って高速処理を実現する。

【００２６】以上の実施形態において、特定話者の発声音声データ３０と、ＨＭ網１１と、ＬＲテーブル１３と、文脈自由文法データベース２０とはそれぞれ、例えばハードディスクメモリに格納される。また、音素照合部４とＬＲパーザ５と不特定話者モデル作成部３１は例えばデジタル電子計算機によって構成される。

【００２７】以上の実施形態においては、図２の不特定話者モデル作成処理によって不特定話者モデルを作成しているが、当該作成処理によって作成されたＨＭ網に対して公知のバーム・ウェルチの学習アルゴリズムを用いて再学習して、ＨＭ網を作成してもよい。

【００２８】

【実施例】本発明者は、図１の音声認識装置の有効性を確かめるために、以下の通り実験を行った。当該実験には、コンテキスト依存型の音素ＨＭＭの状態を効果的に共有したＨＭ網（例えば、従来文献３「鷹見ほか，“音素コンテキストと時間に関する逐次状態分割による隠れマルコフ網の自動生成”，電子通信情報学会技術研究報告，ＳＰ９１－８８，１９９１年１２月」参照。）を使用した。ＨＭ網の構造は１人の話者の発声した２６２０単語の音声データを用いて決定し、総状態数２００、及び６００の２種類のモデルを作成した。各モデルには１状態１０混合の無音モデルを付加した。特定話者モデル学習用の初期話者モデルは無音モデルを除き各状態とも単一分布としパラメータの初期値は構造決定と同じ音声データで決定した。この初期話者モデルをもとに、本特許出願人が所有する、トラベル・ブランニングをタスクとした自然発話の音声認識データベース（例えば、従来文献４「Ｔ．Ｍｏｒｉｍｏｔｏ ｅｔ ａｌ．，“Ａ Ｓｐｅｅｃｈ ａｎｄＬａｎｇｕａｇｅ Ｄａｔａｂａｓｅ ｆｏｒ Ｓｐｅｅｃｈ Ｔｒａｎｓｌａｔｉｏｎ Ｒｅｓｅａｒｃｈ”，Ｐｒｏｃ．ｏｆ ＩＣＳＬＰ’９４，ｐｐ．１７９１－１７９４，１９９４年」参照）の中の男性８１名の自然発話データを用いて最尤推定法により出力ガウス分布の平均値と分散を学習することにより８１名分の特定話者モデルを作成した。但し、１人あたりのデータ量が２０発話程度と少ないため、分散は初期パラメータより値が大きくなる場合のみ更新した。なお、今回は男性話者のみを用いて不特性話者モデルの作成、及び認識実験を行なった。認識実験は学習に用いたものと同じ自然発話データベースより選択した学習データに含まれない男性９人に対して行なった。

【００２９】不特定話者モデルはＨＭ網全体を単位としたモデルベースのクラスタリングを用いた第２の従来例のＣＣＬ法と本発明に係るＨＭＭの状態別クラスタリングの結果を用いる方法により作成し両者の性能を音素認識実験により比較した。ただし、本発明に係る状態別クラスタリングによる方法では特定話者モデルの各状態の

出力ガウス分布の内、学習時の状態占有データ量が１０フレーム以上のもののみを使用した。さらに、状態別クラスタリングによって作成したモデルを初期モデルとしてバーム・ウェルチの学習アルゴリズムによって再学習したモデルの認識率との比較も行なった。またさらに、本発明に係る状態別クラスタリングによる方法でＨＭＭ＊を作成した後、バーム・ウェルチの学習アルゴリズムによって再学習したモデルの認識率についても実験を行った。ここで、実験条件である、分析条件、使用パラメータ、学習／認識データを表１に示す。

【００３０】

【表１】

実験条件

分析条件	サンプリング周波数＝１２ＫＨｚ ハミング窓＝２０ｍｓ フレーム周期＝５ｍｓ
使用パラメータ	１６次ＬＰＣケプストラム＋１６次Δケプストラム ＋対数パワー＋Δ対数パワー
学習データ	男性８１名－－各話者１会話（合計１７９９発声）
不特定話者モデル評価データ	男性９名－－各話者１会話（１１～２９発声）

【００３１】表２及び表３に、第２の従来例のＣＣＬ法（以下、表においてモデルクラスタリングと略す。）及び、本発明に係る状態別クラスタリングによる方法（以下、表において、状態別クラスタリングと略す。）で作成した各状態、混合数のＨＭ網に含まれる出力ガウス分布の総数を示す。第２の従来例のＣＣＬ法による場合は無音モデルを除き全ての状態に対して混合分布数が等しくなるが、本発明に係る状態別クラスタリングによる場合は各状態に対して特定話者モデルから抽出された１０フレーム以上のデータで学習された出力ガウス分布数が設定した混合数より少ない場合には抽出された分布数が※その状態の混合分布数となるためモデルベースのクラスタリングによる場合より総分布数が少なくなっている。但し、今回は各状態における抽出した出力ガウス分布の平均値のばらつきの度合は混合数の決定において考慮していない。このように音素バランスを考慮した音声データの収集が困難な自由発話音声データベースを用いた場合には各状態ごとに混合分布数を設計することにより不要なパラメータの増加を防ぐことができる可能性があることがわかる。

【００３２】

【表２】

不特定話者モデルの総分布数－２０１状態のＨＭ網の場合

作成法／混合数	5	１０	１５	２０
モデルクラスタリング	１０１０	２０１０	３０１０	４０１０
状態別クラスタリング	９７９	１９０３	２７９８	３６７８

【００３３】

★40★【表３】

不特定話者モデルの総分布数－６０１状態のＨＭ網の場合

作成法／混合数	3	5	１０	１５
モデルクラスタリング	１８１０	３０１０	６０１０	９０１０
状態別クラスタリング	１６１７	２５４０	４６１４	６４４７

【００３４】表４及び表５は各方法により作成した不特定話者モデルを用いた音素認識実験の結果である。表中の結果は男性９人に対する平均値を示している。

【００３５】

【表4】

モデル作成法による音素認識率（%）の比較－201状態のHM網の場合

作成法／混合数	5	10	15	20
バーム・ウェルチ	65.9	66.8	－	－
モデルクラスタリング	62.2	62.5	63.3	63.2
状態別クラスタリング	63.6	64.1	64.0	64.5
状態別クラスタリング＋バーム・ウェルチ	68.0	68.6	－	－

【0036】

＊　＊【表5】

モデル作成法による音素認識率（%）の比較－601状態のHM網の場合

作成法／混合数	3	5	10	15
バーム・ウェルチ	67.6	67.8	－	－
モデルクラスタリング	65.1	65.5	66.2	66.2
状態別クラスタリング	67.8	67.9	67.8	67.8
状態別クラスタリング＋バーム・ウェルチ	69.2	69.2	－	－

【0037】表4及び表5の結果を表2及び表3の結果とあわせて見ると、本発明に係る状態別クラスタリングによる方法は全ての条件のもとで第2の従来例のCCL法による場合より少ないパラメータ数で高い認識性能を示しており、認識率の差はHM網の状態数が201状態の場合より601状態の場合の方が大きくなっている。実際の認識処理のスピードや話者適応を行なう場合の効率を考えた場合できるだけ少ないパラメータ数で高い認識性能が得られる方が不特定話者モデルとしての性能は良いと考えられ、このことは、本発明に係る状態別クラスタリングによる方法が性能の良いモデルを得るのに有効な方法であることを示している。

【0038】また、HM網の状態数と認識性能の関係を見た場合、601状態のHM網は201状態のHM網より高い認識性能を示しており、これは、第2の従来例のCCL法及び、本発明に係る状態別クラスタリング法のどちらの場合にも同様のことが言える。これは、201状態ではまだ音韻環境が十分に細分化されてモデル化されていないことが原因であると考えられる。音韻環境が十分に細分化されるように状態分割されていなければ、各状態の出力ガウス分布は音韻環境及び話者環境の両方の要因による音響的特徴量の変動を同時に表現しなければならず、音韻性と話者性の区別が難しくなり、認識誤りの可能性が高くなると考えられる。

【0039】さらに、表4及び表5から明らかなように、本発明に係る状態別クラスタリング法でクラスタリングした後バーム・ウェルチの学習アルゴリズムを用いて再学習した場合、他の方法に比較してより高い音素認識率が得られている。

【0040】最後に、不特定話者モデルの作成時間について述べる。従来文献2において開示された第2の従来例のCCL法では、バーム・ウェルチの学習アルゴリズムの数パーセント程度の計算時間しか要しないと報告されている。本発明に係る状態別クラスタリングを用いる場合にはクラスタリングを行なう回数が増える分、第2の従来例のCCL法に比較して計算時間が増加するが、この時間はモデル作成に要する時間の大部分を占める特定話者モデルの学習時間に比較すると非常に小さいため、全体の時間で見た場合には、第2の従来例のCCL法と同様にバーム・ウェルチの学習アルゴリズムの数パーセント程度の計算時間で不特定話者モデルを作成可能である。

【0041】以上説明したように、本発明に係る実施形態によれば、入力された複数の特定話者の単一ガウス分

布のＨＭＭの各状態の出力ガウス分布を各状態ごとに独立にクラスタリングして合成することにより不特定話者の混合ガウス分布のＨＭＭを作成するので、各特定話者モデルの全てのパラメータが学習されている必要はなく、また話者ごとに学習されているパラメータが異なっていてる場合にも対応することができる。従って、発話数が少ない話者の音声データや自由発話音声のような話者ごとに発話内容が異なるデータに対しても使用することができる。さらに、ＨＭＭの状態ごとに各特定話者モデルから取り出された出力ガウス分布の平均値のばらつきやその学習データ量の情報を利用することによって状態ごとに分割するクラスタ数を決めることができるため、学習データ量や話者間の音響的特徴の変動の度合を考慮した混合分布数をＨＭＭの各状態ごとに決定することができる。当該不特定話者モデルのＨＭＭを用いて音声認識することにより、従来例に比較して高い音声認識率で音声認識することができる。

【００４２】

【発明の効果】以上詳述したように本発明に係る請求項１記載の不特定話者モデル作成装置によれば、入力された複数の特定話者の単一ガウス分布の隠れマルコフモデルに基づいて、不特定話者の混合ガウス分布の隠れマルコフモデルを作成する不特定話者モデル作成装置において、入力された複数の特定話者の単一ガウス分布の隠れマルコフモデルの各状態の出力ガウス分布を各状態ごとに独立にクラスタリングして合成することにより不特定話者の混合ガウス分布の隠れマルコフモデルを作成するモデル作成手段を備える。具体的には、上記モデル作成手段は、入力された複数の特定話者の発声音声データに基づいて、複数の話者に対して同一の初期話者隠れマルコフモデルを用いて所定の学習法により上記発声音声データの存在する状態に対してのみ出力ガウス分布を学習することにより、複数個の特定話者用単一ガウス分布の隠れマルコフモデルを作成する学習手段と、上記学習手段によって作成された複数個の特定話者用単一ガウス分布の隠れマルコフモデルに基づいて、各出力ガウス分布間の距離を基準にして、各クラスタにより短い距離に出力ガウス分布が含まれるように複数のクラスタにクラスタリングを行うクラスタリング手段と、上記クラスタリング手段によって各状態毎にクラスタリングされた単一ガウス分布の隠れマルコフモデルに基づいて、各クラスタ内の複数の出力ガウス分布の隠れマルコフモデルを各状態の単一ガウス分布の隠れマルコフモデルに合成する合成手段と、上記合成手段によって合成された各状態の単一ガウス分布の隠れマルコフモデルを混合することにより、不特定話者の混合ガウス分布の隠れマルコフモデルを作成する混合手段とを備える。

【００４３】すなわち、多数の特定話者モデルから学習されている出力ガウス分布のみを取り出してＨＭＭの各状態で独立にクラスタリングを行なうことにより、各状態における特徴量の変動の大きさや学習データ量を考慮してクラスタ数を決定することが可能となり各状態ごとに最適な出力ガウス分布数を決定することができる。また、各特定話者モデルの学習されている出力ガウス分布のみを選択的に使用することができるため各特定話者モデルの全ての出力ガウス分布が学習されている必要はなく、一人あたりの発話量の少ないデータベースに対しても有効に使用することができる。また、各話者ごとに別々にパラメータ推定を行なうため、全てのデータを一度に使って学習する第１の従来例のバーム・ウェルチの学習アルゴリズムによる方法に対して計算量を飛躍的に減らすことが可能となる。従って、不特定話者モデルの作成時間を大幅に短縮することができる。

【００４４】また、請求項３記載の不特定話者モデル作成装置によれば、上記クラスタリング手段は、各状態毎に予め設定したしきい値以上のデータ量で学習された出力ガウス分布のみを取り出した後、クラスタリングする。これにより、信頼性のより高い最適な不特定話者モデルを作成することができる。従って、当該不特定話者モデルを用いて音声認識を行うことにより、従来例に比較してより高い音声認識率で音声認識することができる。

【００４５】さらに、請求項４記載の不特定話者モデル作成装置によれば、上記クラスタリング手段は、各状態においてクラスタリングされた各クラスタの中心と各出力ガウス分布間の距離の平均値が予め決めた距離以下になるまでクラスタリングを繰り返すことにより、各状態における各出力ガウス分布のバラツキが大きいほどクラスタ数が多くなるように各状態におけるクラスタ数を決定する。従って、各状態における各出力ガウス分布のバラツキを考慮してクラスタ数を決定することが可能となり各状態ごとに最適な出力ガウス分布数を決定することができる。これにより、信頼性のより高い最適な不特定話者モデルを作成することができる。それ故、当該不特定話者モデルを用いて音声認識を行うことにより、従来例に比較してより高い音声認識率で音声認識することができる。

【００４６】また、本発明に係る請求項５記載の音声認識装置によれば、入力された複数の特定話者の単一ガウス分布の隠れマルコフモデルに基づいて、不特定話者の混合ガウス分布の隠れマルコフモデルを作成する請求項１乃至４のうちの１つに記載の不特定話者モデル作成装置と、入力された発声音声文の音声信号に基づいて、上記不特定話者モデル作成装置によって作成された不特定話者の混合分布の隠れマルコフモデルを用いて、音声認識する音声認識手段とを備える。従って、当該不特定話者モデルを用いて音声認識を行うことにより、従来例に比較してより高い音声認識率で音声認識することができる。

【図面の簡単な説明】

15　　16

【図１】　本発明に係る一実施形態である音声認識装置のブロック図である。

【図２】　図１の不特定話者モデル作成部によって実行される不特定話者モデル作成処理を示すフローチャートである。

【図３】　図１の不特定話者モデル作成部によって実行される不特定話者モデル作成処理のうち特定話者モデルの学習と出力ガウス分布の抽出の処理を示す図である。

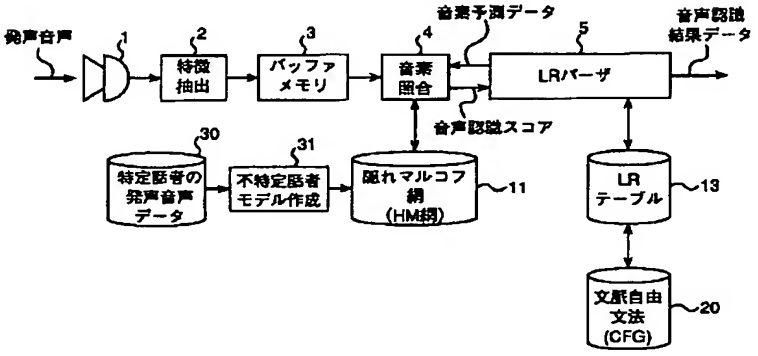
【図４】　図１の不特定話者モデル作成部によって実行される不特定話者モデル作成処理のうち各状態毎の出力ガウス分布のクラスタリングの処理を示す図である。

【図５】　図１の不特定話者モデル作成部によって実行される不特定話者モデル作成処理のうち各クラスタ毎に＊＊複数の確率密度関数を混合する処理を示す図である。

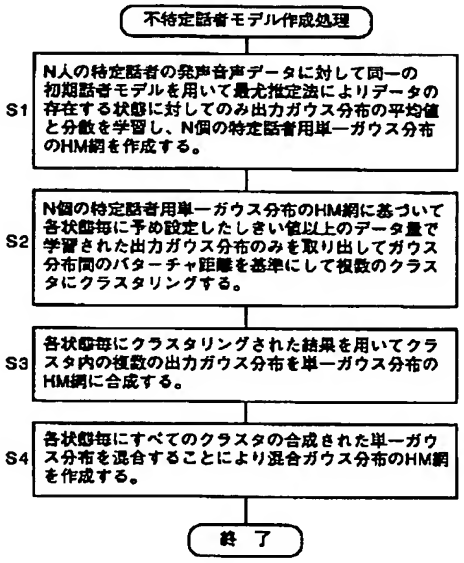
【符号の説明】

１…マイクロホン、
２…特徴抽出部、
３…バッファメモリ、
４…音素照合部、
５…ＬＲパーザ、
１１…隠れマルコフ網（ＨＭ網）、
１３…ＬＲテーブル、
２０…文脈自由文法データベース、
３０…特定話者の発声音声データ、
３１…不特定話者モデル作成部。

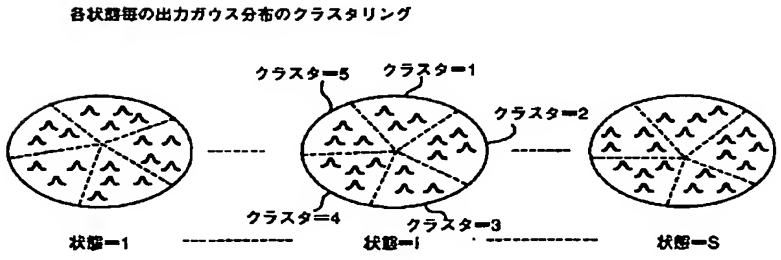
【図１】



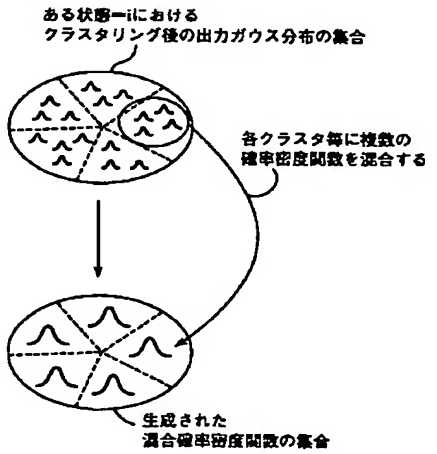
【図２】



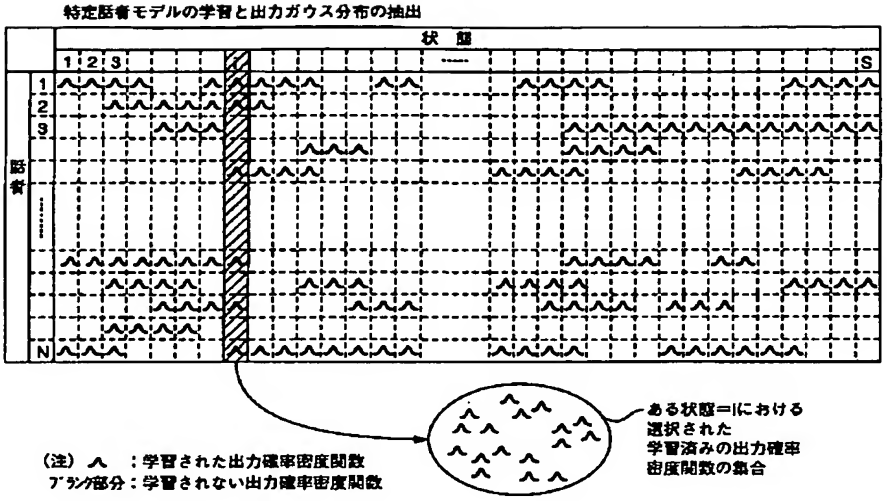
【図４】



【図５】



【図３】


特定話者モデルの学習と出力ガウス分布の抽出
状態
1 2 3 I S
話者
1
2
3
N
ある状態=Iにおける
選択された
学習済みの出力確率
密度関数の集合
(注) ∧ ：学習された出力確率密度関数
ﾌﾟﾗﾝｸ部分：学習されない出力確率密度関数